Trends in Skills Set Requirement among Software Developers

George Mathew

16th May 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

Disclaimer: As the data used for the project is a part (1/6th) of the survey data, hence the results may vary from the original results of the survey available in https://insights.stackoverflow.com/survey/2020

EXECUTIVE SUMMARY

- My organization regularly analyses data to help identify future skill requirements. For that we collect the top technology skills that are most in demand from various sources including:
    - Job postings
    - Training portals
    - Surveys
- The prepared data following data wrangling is analysed using statistical techniques and visualization of charts and interactive dashboards. The exercise is undertaken keeping in mind our objective to find out the answers to the following problems.
    - What are the top programming languages that are in demand ?
    - What are the top database skills that are in demand ?
    - What are the most popular Platforms ?
    - Demographic data like gender and age distribution of developers.

EXECUTIVE SUMMARY

- The answers that were found to the problems based on the exercise are :
  - The top ten programming languages in descending order are : JavaScript, HTML/CSS, TypeScript, SQL, Python, C#, Go, Bash/Shell/PowerShell, Rust, Java
  - The top ten databases in descending order are : PostgreSQL, Redis, MongoDB, Elasticsearch, MySQL, Microsoft SQL Server, SQLite, Firebase, MariaDB, DynamoDB
  - The top ten Platforms in descending order are : Docker, Linux, AWS, Kubernetes, Windows, Microsoft Azure, Google Cloud Platform, MacOS, Android, iOS.
  - The demographic data based on the survey shows that almost 90% of the respondents in the survey were men and the maximum developers were in the age group 25 to 35.

- The exercise has resulted in identifying the future skill requirements which will help :
  - The developers add the required skill set to their existing skills and remain relevant in the competition
  - The employers to guide and support their existing developers to reskill themselves as also to align the training infrastructure of the employers and training organisations.

INTRODUCTION

- I have joined as a Data Analyst in a global IT and business consulting services firm that is known for their expertise in IT solutions and has a team of highly experienced IT consultants. To keep pace with changing technologies and remain competitive, my organization regularly analyzes data to help identify future skill requirements.
- The business challenge before me requires collecting data and performing data analysis on real-world datasets identifying trends for this year's report on emerging skills
- The various sources for collecting data include :

  >> Job postings >> Training portals >> Surveys
- Prepare the collected data for analysis by using data wrangling techniques like, finding duplicates, removing duplicates, finding missing values, and inputting missing values.
- Apply statistical techniques to analyze the data and identify insights and trends.
- Choose an appropriate visualization based on the data and make a presentation using charts, plots, and histograms to help reveal the findings and trends and use interactive dashboards to help analyze and present the data dynamically with analysis report and an executive summary to the various stakeholders in the organization.
- This report will come handy to the developers as the identification of future skill requirements which will help them add the required skill set to their existing skills and remain relevant in the competition.
- The report will act as a guide to employers / organisations to support their existing developers to reskill themselves and align the training infrastructure of the employers and training organisations.

METHODOLOGY

- We will begin by scraping internet web sites and accessing APIs to collect data in various formats like .csv files, excel sheets, and databases. The source of data is the stackoverflow survey of 2020 available in https://insights.stackoverflow.com/survey/2020.
- Once this is completed, the next step is to get the data ready for analysis using data wrangling techniques.
- As the quest of the organisation is to identify the future skill requirements, the problems for which we need to find answers are :
    - What are the top programming languages that are in demand ?
    - What are the top database skills that are in demand ?
    - What are the most popular Platforms ?
    - Demographic data like gender and age distribution of developers.
- So, when the data is ready, we will apply statistical techniques to analyze the data.
- Then bring all the information together by using charts, plots, histograms as also use IBM Cognos Analytics to create the dashboard and prepare the presentation.

RESULTS

The results of various statistical and visualisation methods used are listed as under:

Data wrangling – Removing duplicates / Removing or replacing Null values /Normalising compensation

- The Language names are classified as Difficult, Hard and Easy with 'Go' being the difficult, 'C#', 'C++' and 'R' being Hard while 'Python', 'Java', 'Javascript', 'Swift' 'PHP' and 'SQL' fall under the Easy category. (Appendix – Fig -1)
- The lowest average salary is for those with skills in PHP and SQL and the highest is for Swift, followed by Python, C++, JavaScript, Java, Go, R and C# in the descending order. (Appendix – Fig -1)

Exploratory Data Analysis – Summary of charts plotted and why used

- The frequency distribution chart shows a normal distribution curve for salary based on the language skills with a low sigma value depicting narrow distribution as far as the salary for different language skills is concerned. (Appendix – Fig - 2)
- The Histogram shows that while one bin has a frequency of 9600 (approx.) other ten bins have a very low frequency ranging upto 200. (Appendix – Fig - 3)

RESULTS (contd/-)

Exploratory Data Analysis - Summary of charts plotted and why used (contd/-)

- The median converted compensation in USD is 57745 (Appendix – Fig - 4)
- The gender distribution shows 10480 men and 731 women (Appendix – Fig - 5)
- The median converted compensation in USD for men is 57744 and for women 57708 (Appendix – Fig - 6)
- The mean age is 30.78 (Appendix – Fig - 7)
- The histogram of Age shows maximum (5500) respondents within the age group of 25 to 35 with drastic reduction after the age of 40 (Appendix – Fig - 8)
- The outliers for converted compensation is negligible. (Appendix – Fig - 9)
- The highest correlation of Age is with converted compensation of 0.40 with the next highest correlation being 0.09 of Age with CodeRevHrs (Appendix – Fig - 10)

Data Visualisation - Summary thereof and why used

- There are outliers above Q3 (towards the last quartile) as far as column Age is concerned (Appendix – Fig - 11)

RESULTS (contd/-)

Data Visualisation - Summary thereof and why used

- The scatter plot of Age and WorkWeekHours makes identification of correlation difficult (Appendix – Fig - 12)
- The bubble plot of Age and WorkWeekHours makes identification of correlation difficult (Appendix – Fig - 13)
- The highest respondents for the survey were from United States (52%) followed by United Kingdom (15%) (Appendix – Fig - 14)
- The maximum respondents (4500 approx.) were developers by profession (Appendix – Fig - 15)

Dashboard Using IBM Cognos Analytics

Results from three interactive dashboards showing trends in the required skill set

Dashboard Tab 1 – Current Technology Usage – (Slide – 17)

- The top programming languages during the current year are JavaScript, HTML/CSS, SQL in the descending order.

RESULTS (contd/-)

Dashboard Using IBM Cognos Analytics

Dashboard Tab 1 – Current Technology Usage (contd/-)

- The top databases during the current year are MySQL, PostgreSQL, Microsoft SQL Server in the descending order
- The top platforms during the current year are Windows, Linux, Docker, AWS, MacOS in the descending order
- The top New Collab Tools during the current year are GitHub, Stack, Google Suite, Jira in the descending order

Dashboard Tab 2 – Future Technology Trend - (Slide – 18)

- The top programming languages during the next year are JavaScript, HTML/CSS, Typescript, SQL in the descending order
- The top databases during the next year are PostgreSQL, Redis, MongoDB in the descending order
- The hierarchy of platforms during the next year are Docker, Linux, AWS, Kubernotes in the descending order

RESULTS (contd/-)

Dashboard Using IBM Cognos Analytics

Dashboard Tab 2 – Future Technology Trend (contd/-)

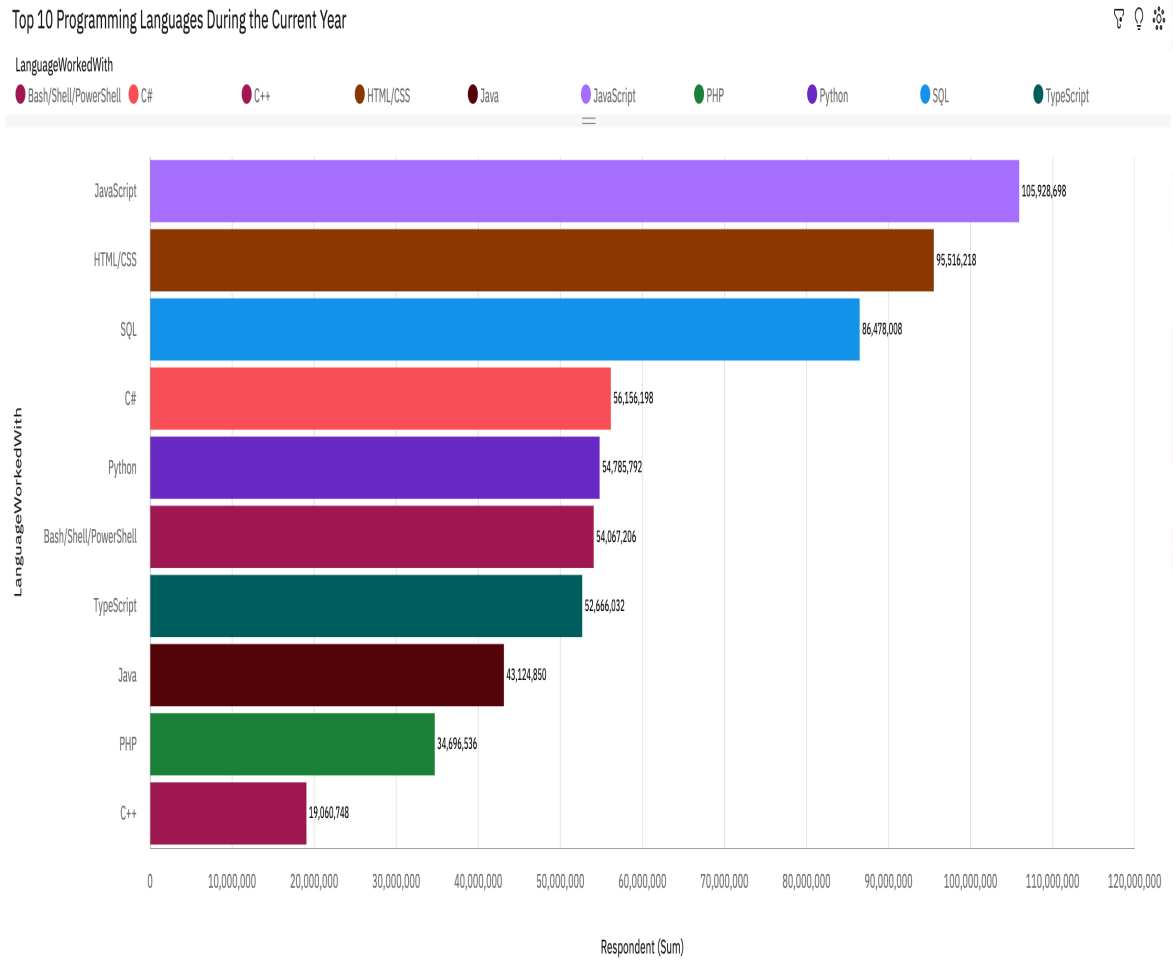
- The top New Collab Tools during the next year are GitHub, Stack, Google Suite, Jira in the descending order

Dashboard Tab 3 – Demographics - (Slide – 19)

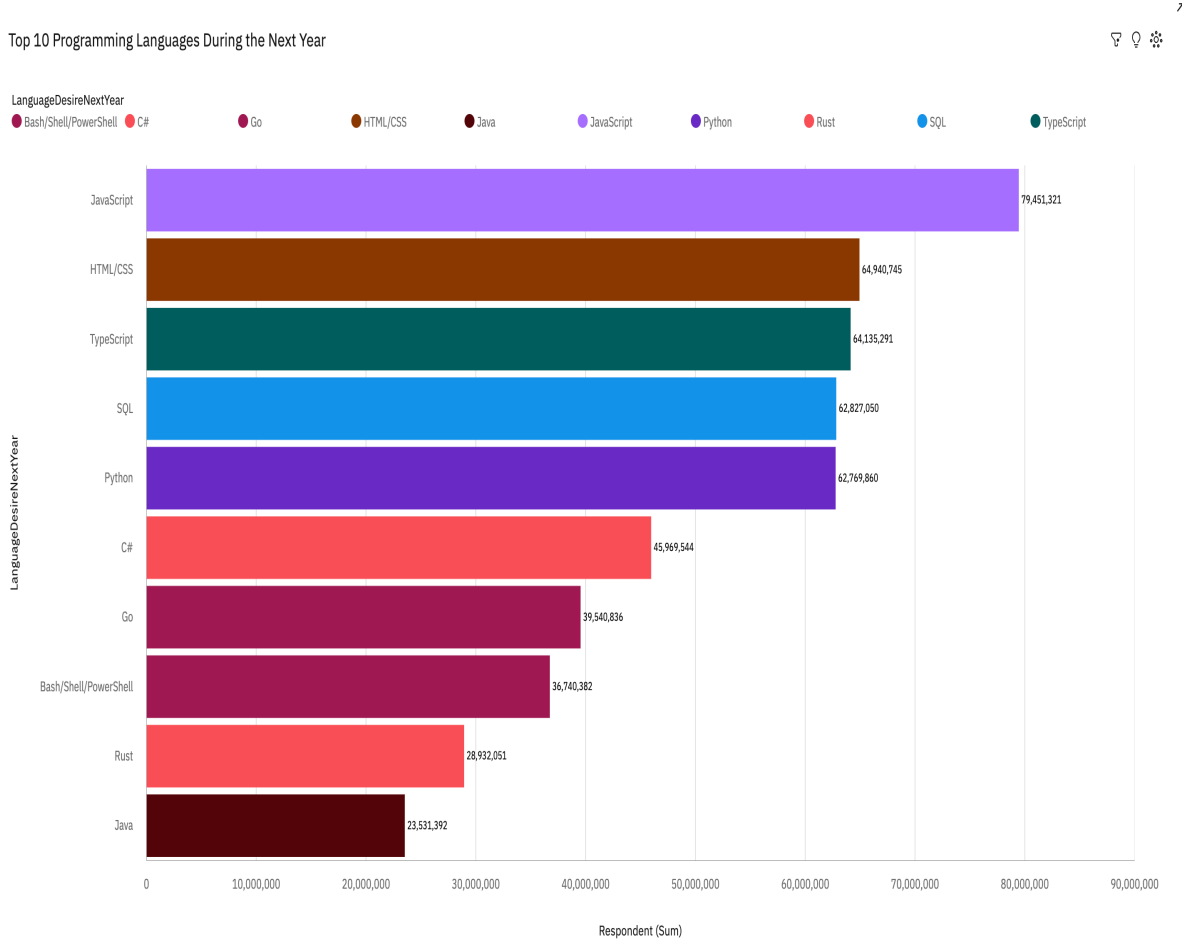
- The distribution of respondents to the survey was 96% men and 4% women
- The maximum respondents to the survey were from United States of America followed by United Kingdom, Germany, India in descending order
- The respondents were mainly in the Age group 22 to 41 and the maximum respondents had an Age of 29
- The education level of maximum of the respondents was graduate followed by masters

PROGRAMMING LANGUAGE TRENDS

Current Year



Next Year



IBM Developer

SKILLS NETWORK

PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

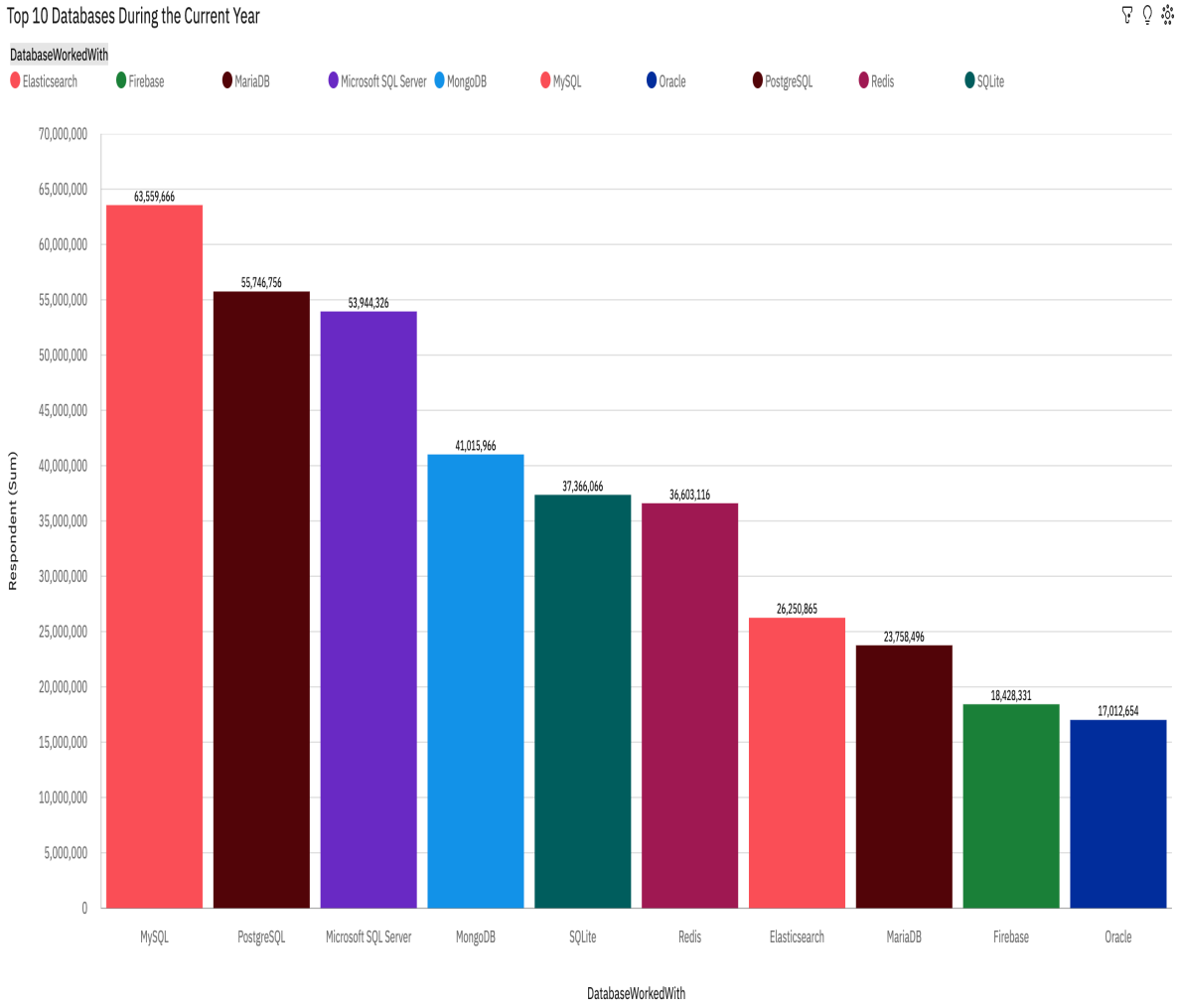
- More respondents have chosen Python and Typescript for next year in comparison to this year.
- The choice of programming languages as priority during the next year has gone down drastically in respect of Bash/Shell/PowerShell, HTML/CSS, Java, JavaScript, SQL.
- C++ and PHP are no more in the top 10 Programming Languages while Go which was not there among the top 10 in the current year has ranked 8th for next year. Rust has also entered the league.

Implications

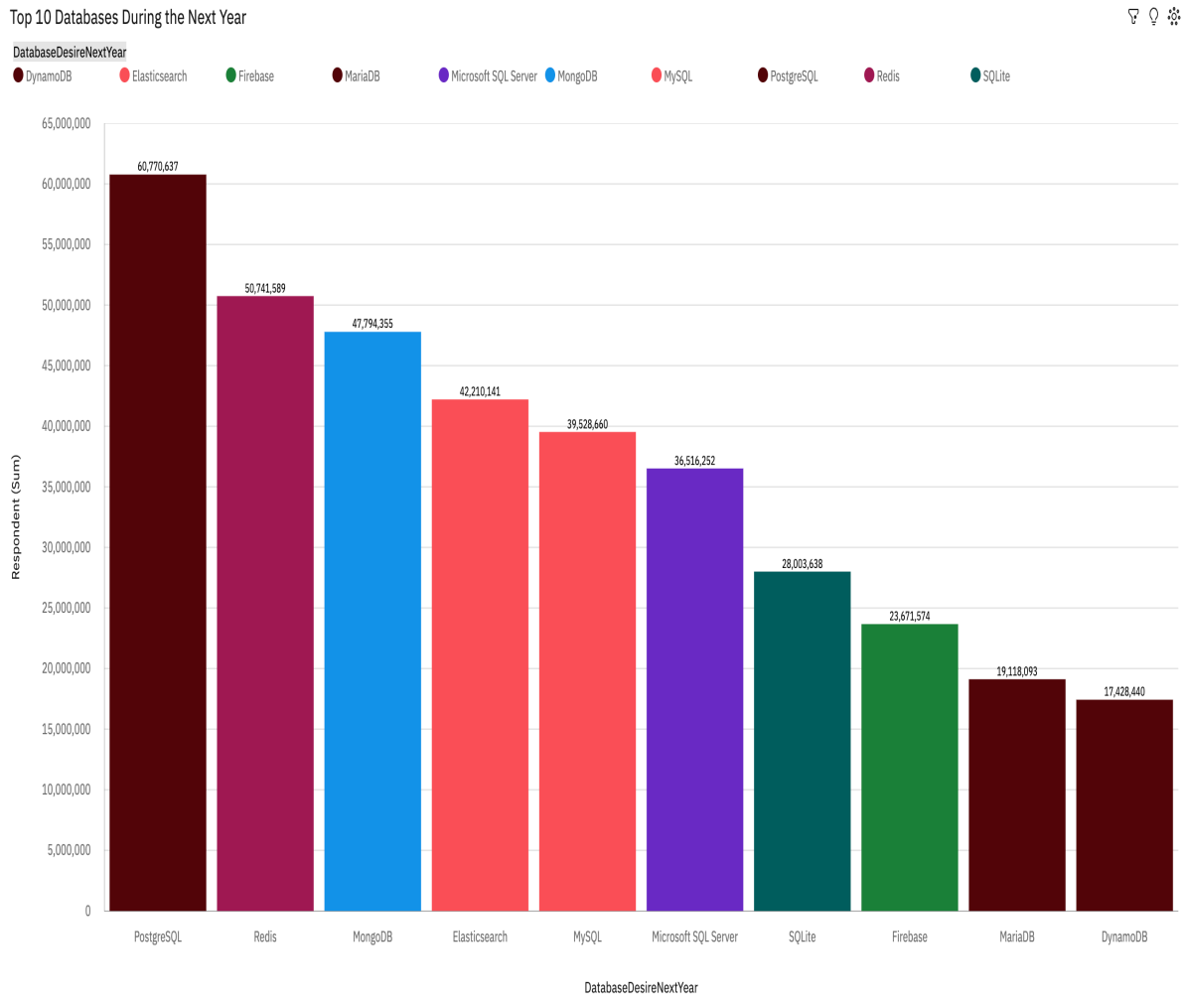
- Demand for both the languages viz. Python and Typescript will increase and may also lead to increase in the compensation for persons with these skills.
- As the demand for these languages reduces the developers will go for skill upgradation by learning other languages whose demand is increasing.
- Languages like Go and Rust will be the choice of many developers and initially the compensation may be high till more developers acquire the skills.

DATABASE TRENDS

Current Year



Next Year



IBM Developer

SKILLS NETWORK

DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

- More respondents have chosen Elasticsearch, Firebase, MongoDB, PostgreSQL and Redis for next year in comparison to this year.
- The choice of database as priority during the next year has gone down drastically in respect of Microsoft SQL Server, MySQL while moderately in case of MariaDB and SQLite.
- Oracle is no more in the top 10 Database trending list while DynamoDB has entered the league.

Implications

- Demand for the Databases viz. Elasticsearch, Firebase, MongoDB, PostgreSQL and Redis will increase and may also lead to increase in the compensation for persons with these skills.
- As the demand for these Databases reduces the developers will go for skill upgradation by learning other Databases whose demand is increasing.
- Databases like DynamoDB will be the choice of many developers and initially the compensation may be high till more developers acquire the skills.

DASHBOARD

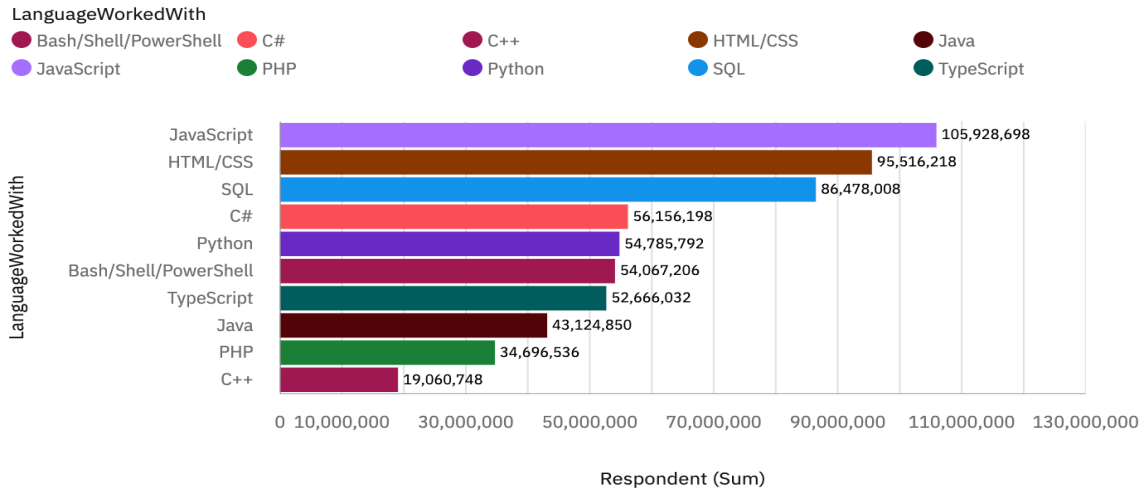
Permanent link of the read-only view of the Cognos dashboard :

https://ap1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2FNew%2Bdashboard_Final%2BAssignment%2BPart%2BI%2BData%2BAnalytics%2Band%2BVisualisation%2BCapstone%2BProject&action=view&mode=dashboard&subView=model000001882054fec1_00000000

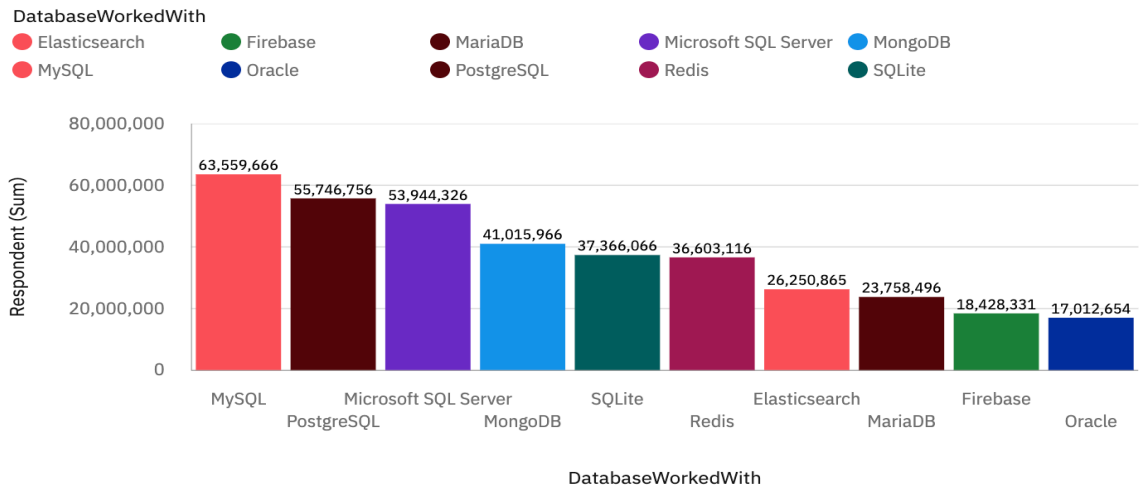
DASHBOARD TAB 1

Current Technology Usage | Future Technology Trend | Demographics

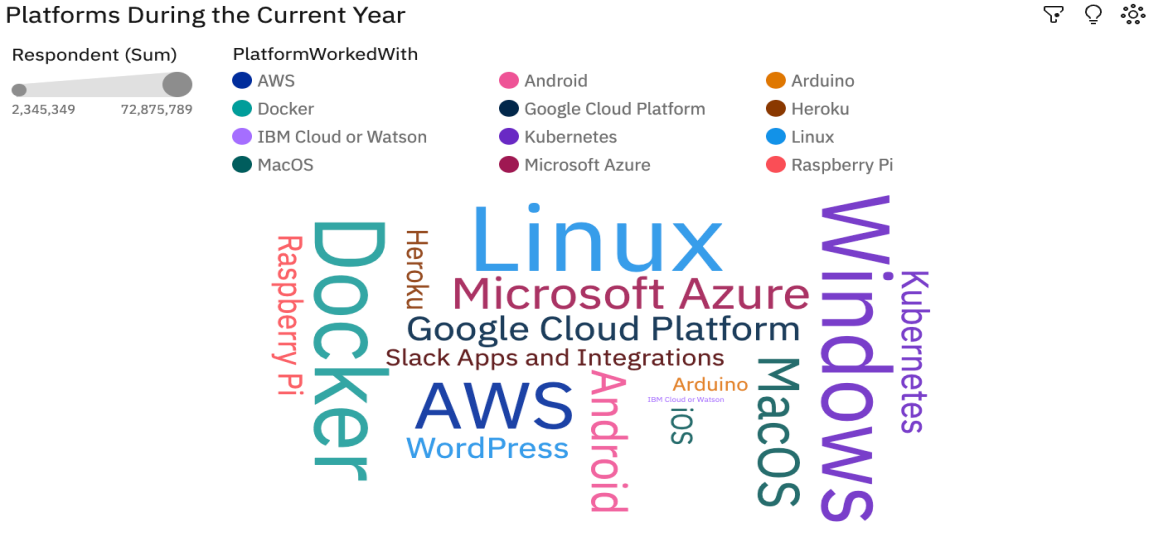
Top 10 Programming Languages During the Current Year



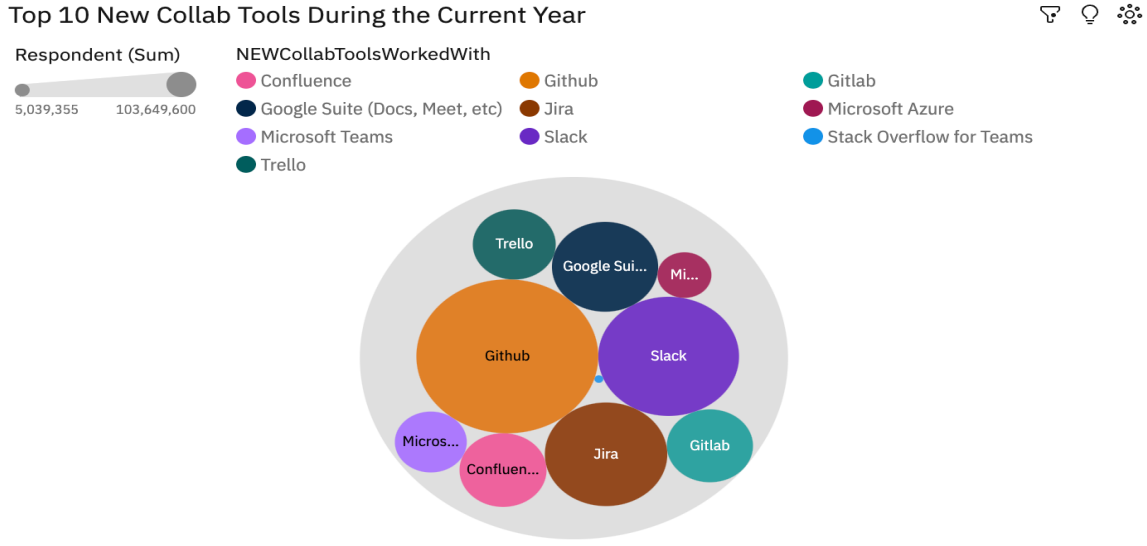
Top 10 Databases During the Current Year



Platforms During the Current Year



Top 10 New Collab Tools During the Current Year



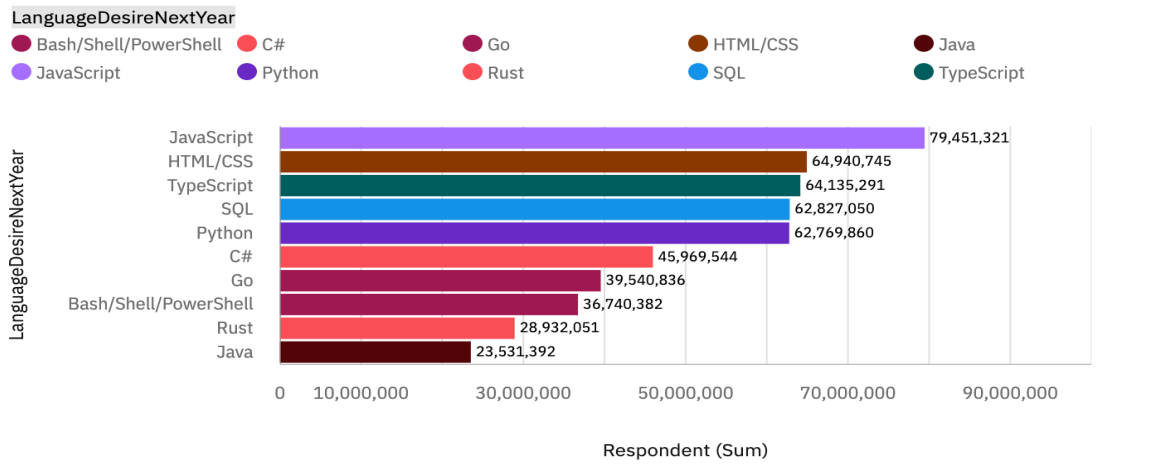
IBM Developer

SKILLS NETWORK

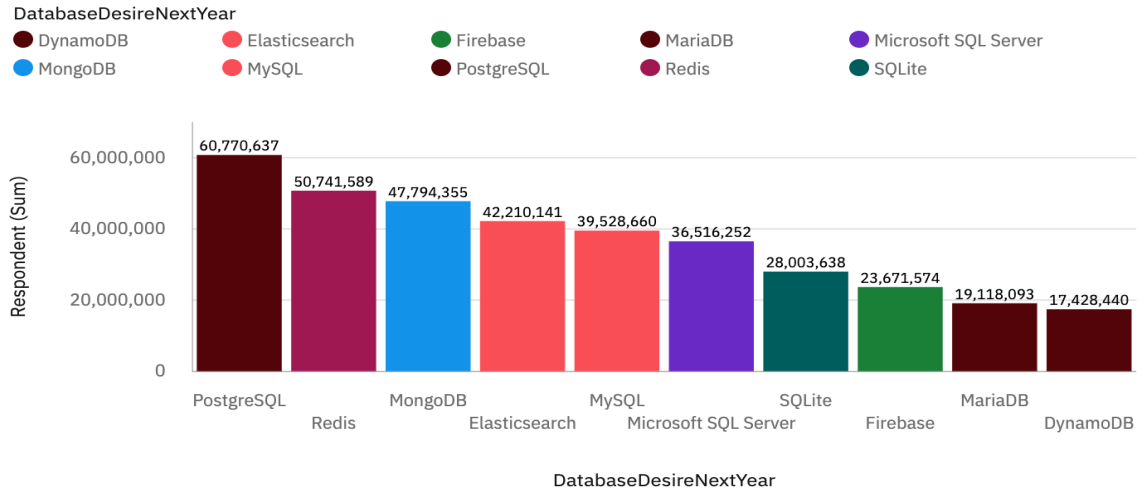
DASHBOARD TAB 2

Current Technology Usage | Future Technology Trend | Demographics

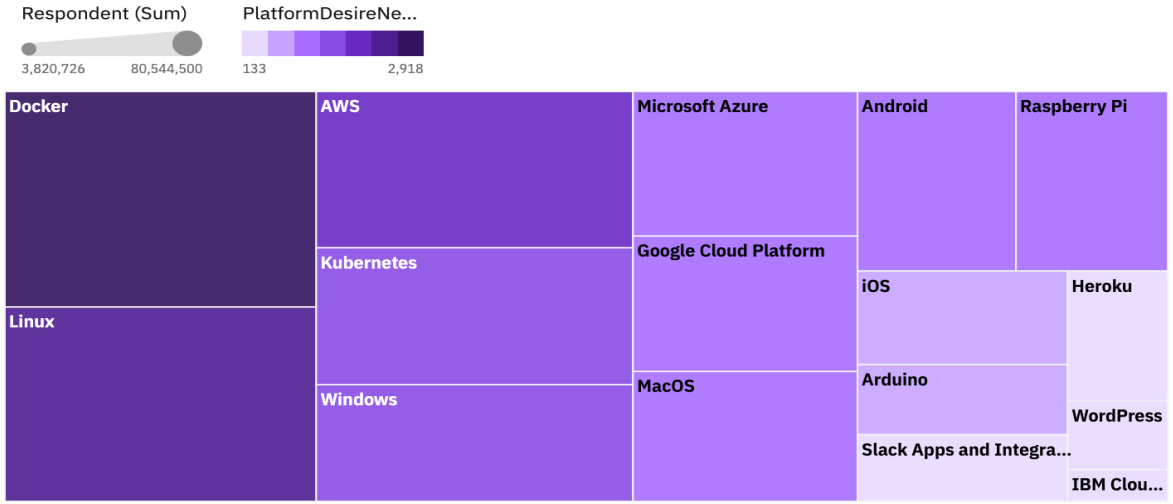
Top 10 Programming Languages During the Next Year



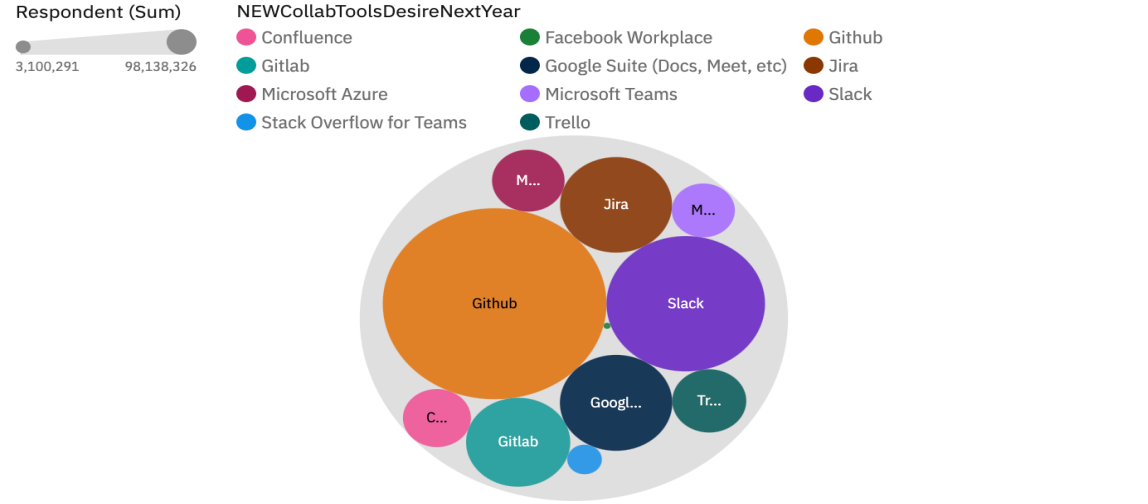
Top 10 Databases During the Next Year



Hierarchy of Platform During the Next Year



Top 10 New Collab Tools During the Next Year



IBM Developer

SKILLS NETWORK

DASHBOARD TAB 3

All tabs

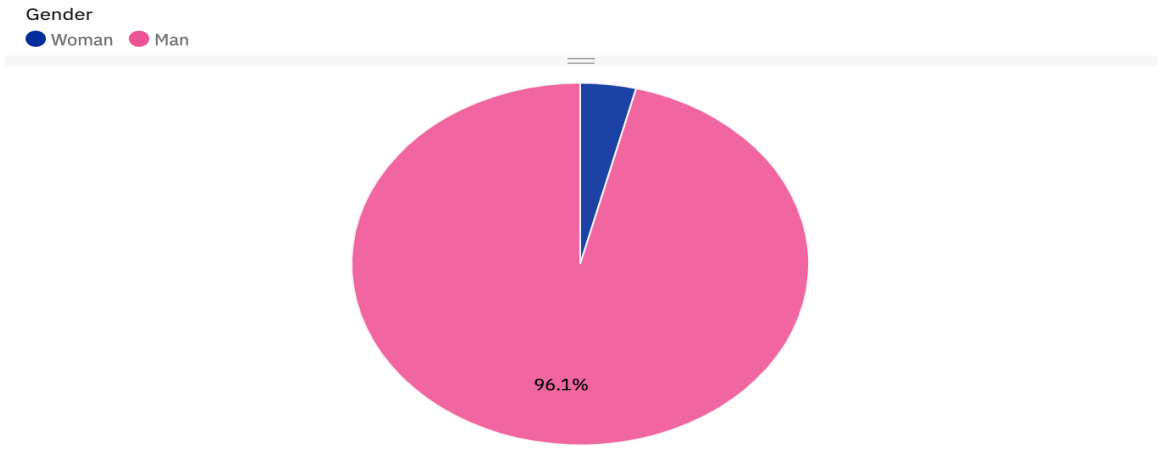
Drag and drop data here to filter all tabs.

This tab

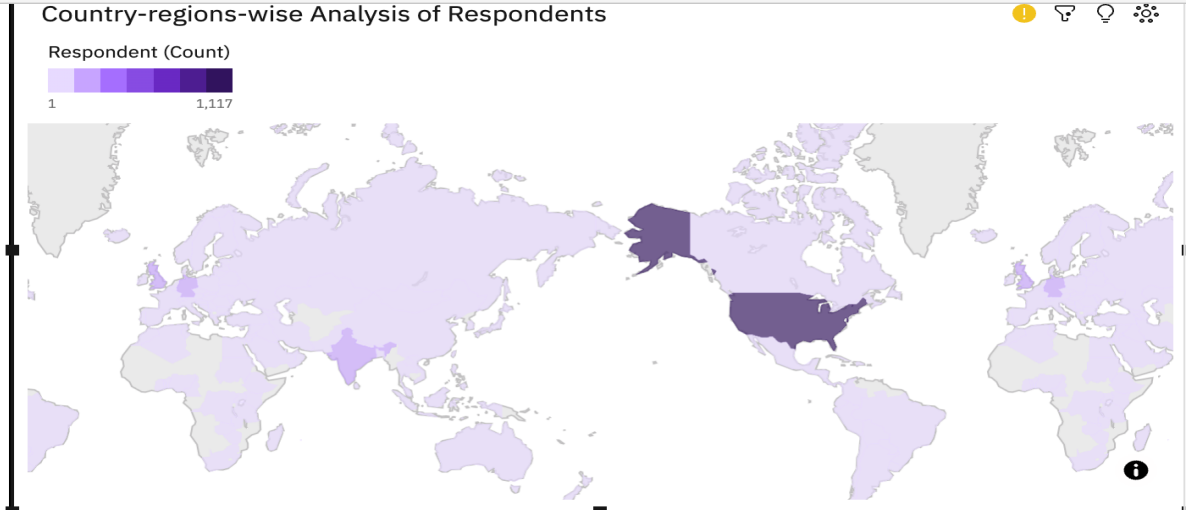
Gender
Man, Woman 2

Current Technology Usage | **Future Technology Trend** | **Demographics** | +

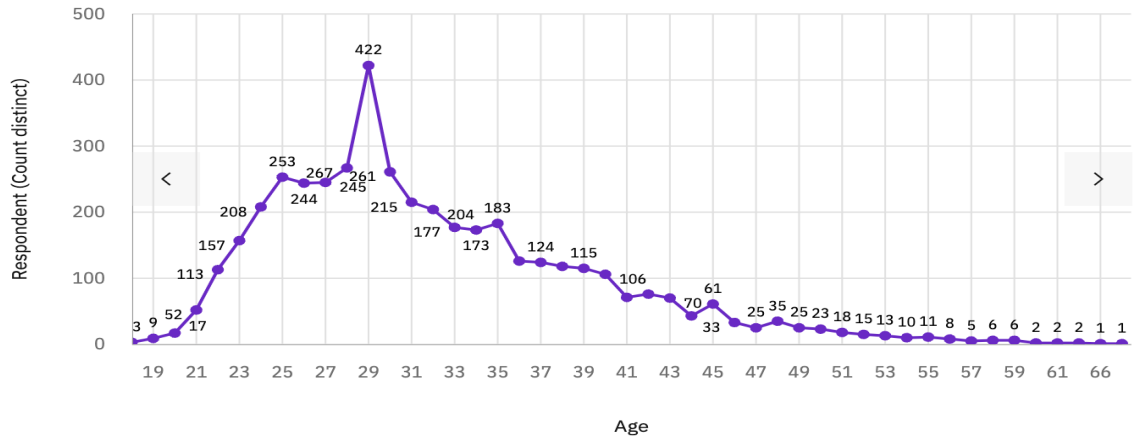
Gender-wise Distribution Percentage



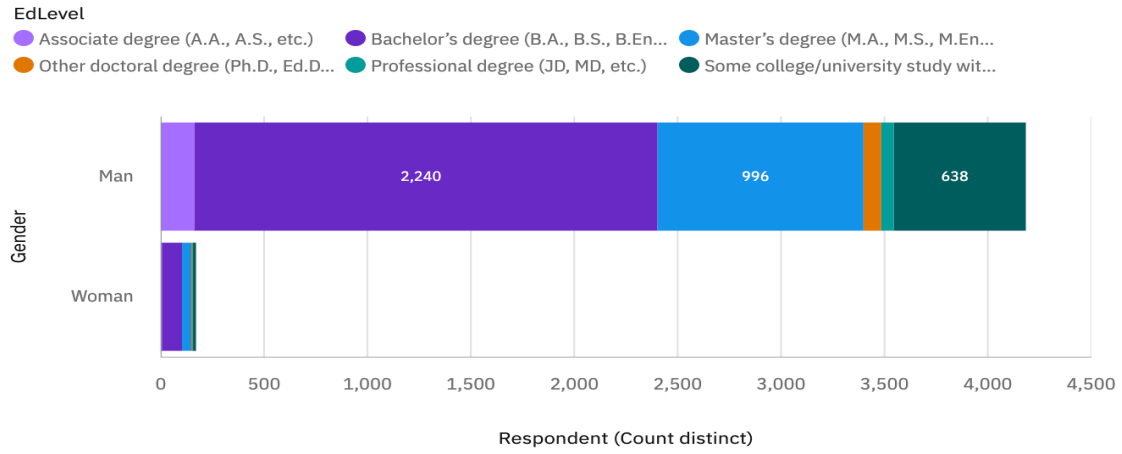
Country-regions-wise Analysis of Respondents



Age-wise Analysis of Respondents



Gender-wise classification on Education Level



DISCUSSION

• The survey data has thrown light on the current and immediate next year scenario in the area of:

>> Programming languages preferred by the developers

>> Databases preferred by the developers

>> Platforms preferred by the developers

>> Collaborative tools preferred by the developers

>> Variation in compensation based on programming language skills

>> Age of the majority group among developers.

OVERALL FINDINGS & IMPLICATIONS

Findings

- Preference for some Programming languages are increasing among developers while for some it is decreasing.
- Preference for some Databases are increasing among developers while for some it is decreasing.
- Correlation between Age and compensation is Medium.
- Maximum developers were in the age group 25 to 35.
- Almost 90% of the respondents in the survey were men.
- It is a normal distribution curve for salary based on the language skills with a narrow distribution.

Implications

- Developers will upgrade their skills in the Programming languages that have increased preferences.
- Developers will upgrade their skills in the Databases that have increased preferences.
- Compensation is not related to Age is a stricter sense.
- Majority of the developers belong to young age group with ability to acquire new skills.
- Either the proportion of men is very high among the developers or women have preferred not to take part in the survey.
- Variation of salary based on the language skills is not large.

CONCLUSION

At the beginning of the project, we defined the Problem for which we want to find answers. The problems are as under :

- What are the top programming languages that are in demand ?
- What are the top database skills that are in demand ?
- What are the most popular Platforms ?
- Demographic data like gender and age distribution of developers.

While answering the first three problems we will answer it from the perspective of future trends as the objective of the whole project is to identify future skill requirements.

CONCLUSION

Problem 1

What are the top programming languages that are in demand ?

Ans : The top ten programming languages in descending order are :

JavaScript, HTML/CSS, TypeScript, SQL, Python, C#, Go, Bash/Shell/PowerShell, Rust, Java

Problem 2

What are the top database skills that are in demand ?

Ans : The top ten databases in descending order are :

PostgreSQL, Redis, MongoDB, Elasticsearch, MySQL, Microsoft SQL Server, SQLite, Firebase, MariaDB, DynamoDB

CONCLUSION

Problem 3

• What are the most popular Platforms ?

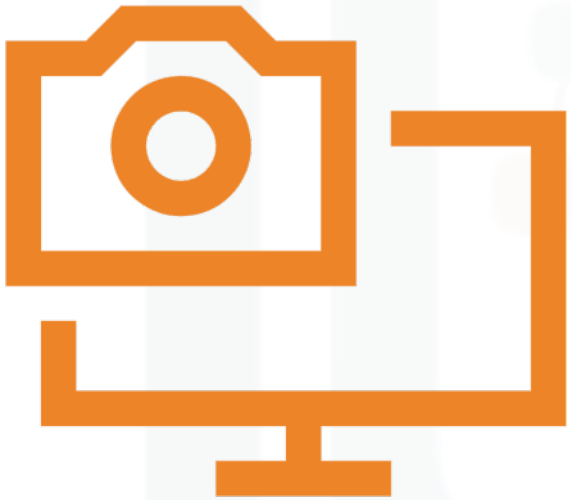
Ans : The top ten Platforms in descending order are :
Docker, Linux, AWS, Kubernetes, Windows, Microsoft Azure, Google Cloud Platform, MacOS, Android, iOS.

Problem 4

• Demographic data like gender and age distribution of developers.

Ans : The demographic data based on the survey shows that almost 90% of the respondents in the survey were men and the maximum developers were in the age group 25 to 35.

APPENDIX

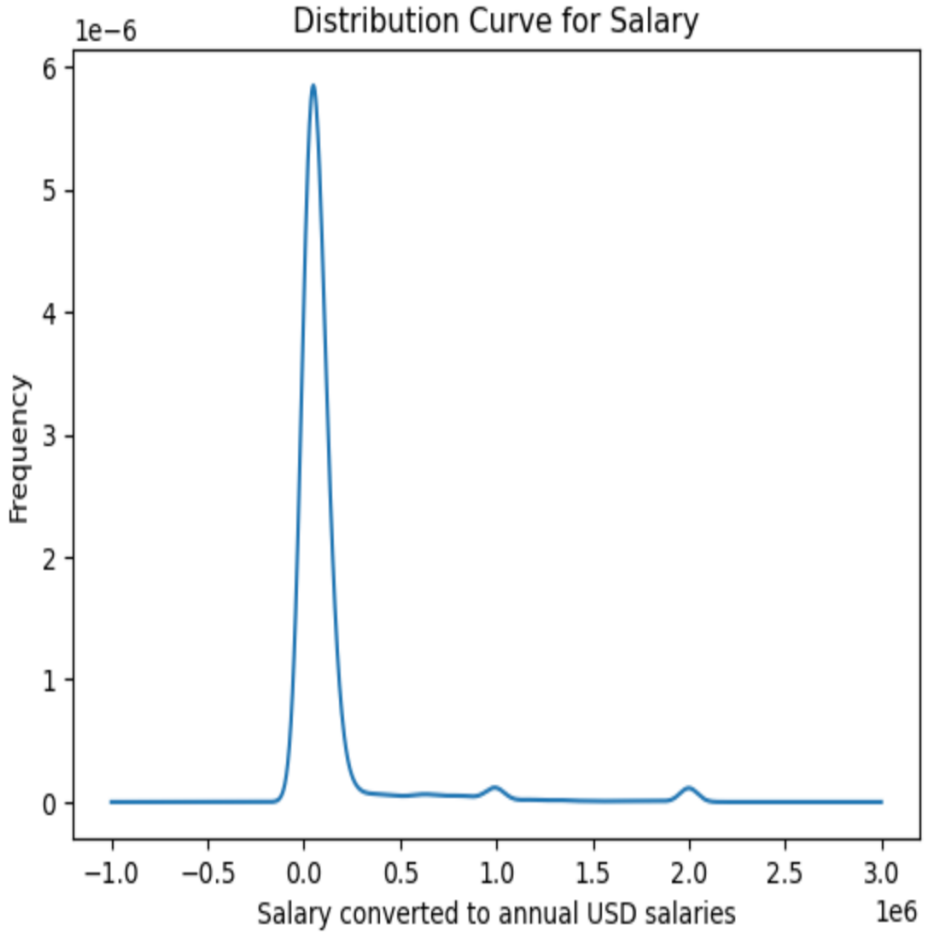


- Include any relevant additional charts, or tables that you may have created during the analysis phase.
- Fig-1

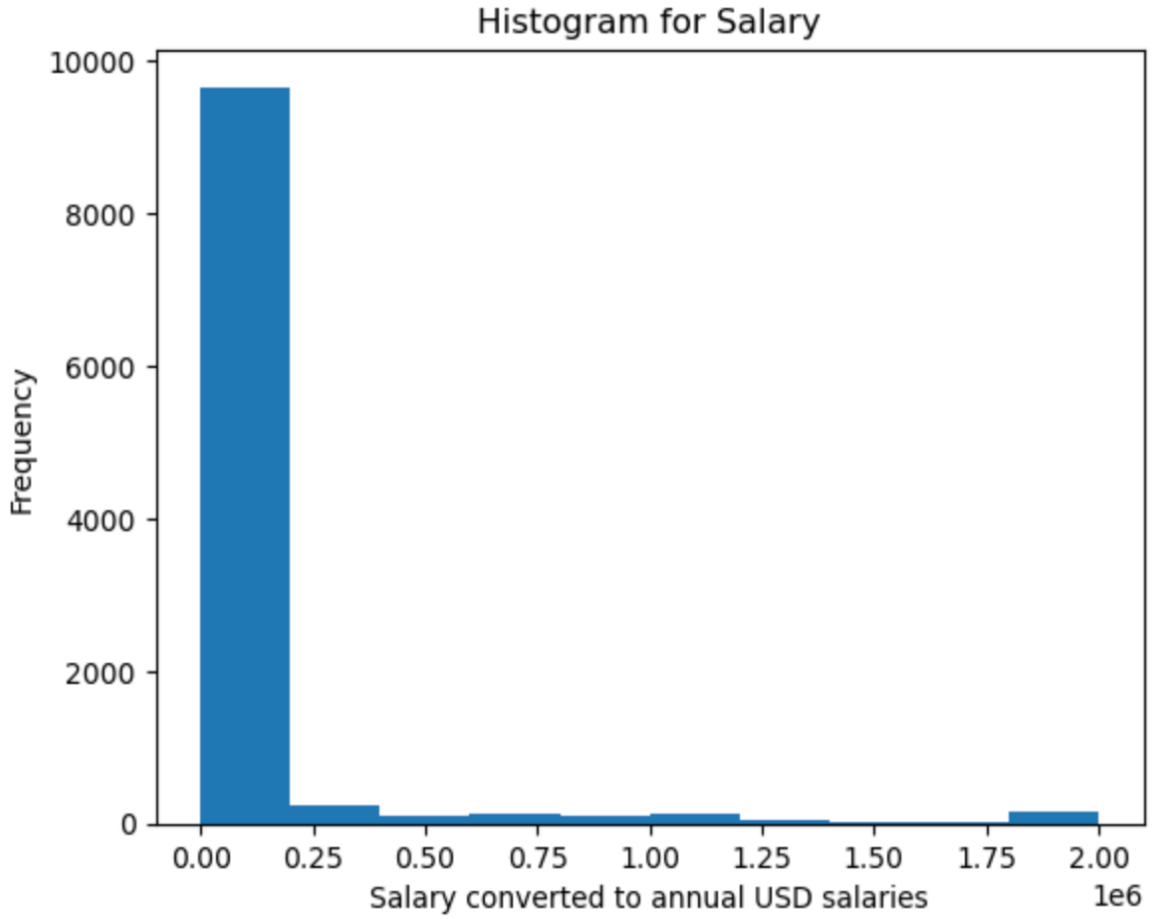
	Language Name	Average Salary	Learning Difficulty
1	Python	$114,383	Easy
2	Java	$101,013	Easy
3	R	$92,037	Hard
4	Javascript	$110,981	Easy
5	Swift	$130,801	Easy
6	C++	$113,865	Hard
7	C#	$88,726	Hard
8	PHP	$84,727	Easy
9	SQL	$84,793	Easy
10	Go	$94,082	Difficult

APPENDIX

- **Fig-2**



- **Fig-3**

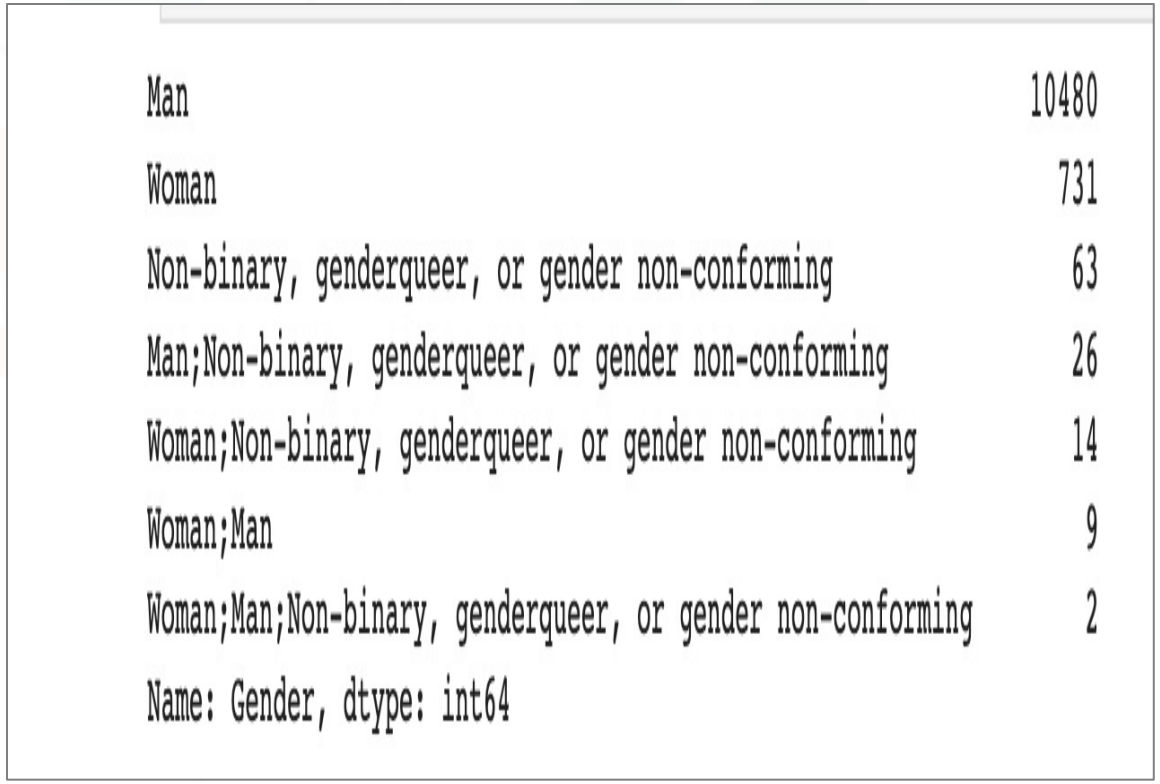


APPENDIX

Fig - 4

- What is the median of the column ConvertedComp ?
- In [711:
- # your code goes here
- df ['ConvertedComp'].median ()
- Out [71]: 57745.0

Fig - 5



```
Man                                                 10480
Woman                                                 731
Non-binary, genderqueer, or gender non-conforming      63
Man;Non-binary, genderqueer, or gender non-conforming  26
Woman;Non-binary, genderqueer, or gender non-conforming 14
Woman;Man                                              9
Woman;Man;Non-binary, genderqueer, or gender non-conforming 2
Name: Gender, dtype: int64
```

APPENDIX

Fig - 6

```
Gender
Man                                                            57744.0
Man;Non-binary, genderqueer, or gender non-conforming          59520.0
Non-binary, genderqueer, or gender non-conforming              67142.0
Woman                                                          57708.0
Woman;Man                                                      21648.0
Woman;Man;Non-binary, genderqueer, or gender non-conforming    30244.0
Woman;Non-binary, genderqueer, or gender non-conforming        65535.5
Name: ConvertedComp, dtype: float64
```

Fig - 7

Give the five number summary for the column `Age` ?

Double click here for hint.

In [74]:
```
# your code goes here
df['Age'].describe()
```

Out[74]:
```
count    11111.000000
mean        30.778895
std          7.393686
min         16.000000
25%         25.000000
50%         29.000000
75%         35.000000
max         99.000000
Name: Age, dtype: float64
```

APPENDIX

Fig - 8

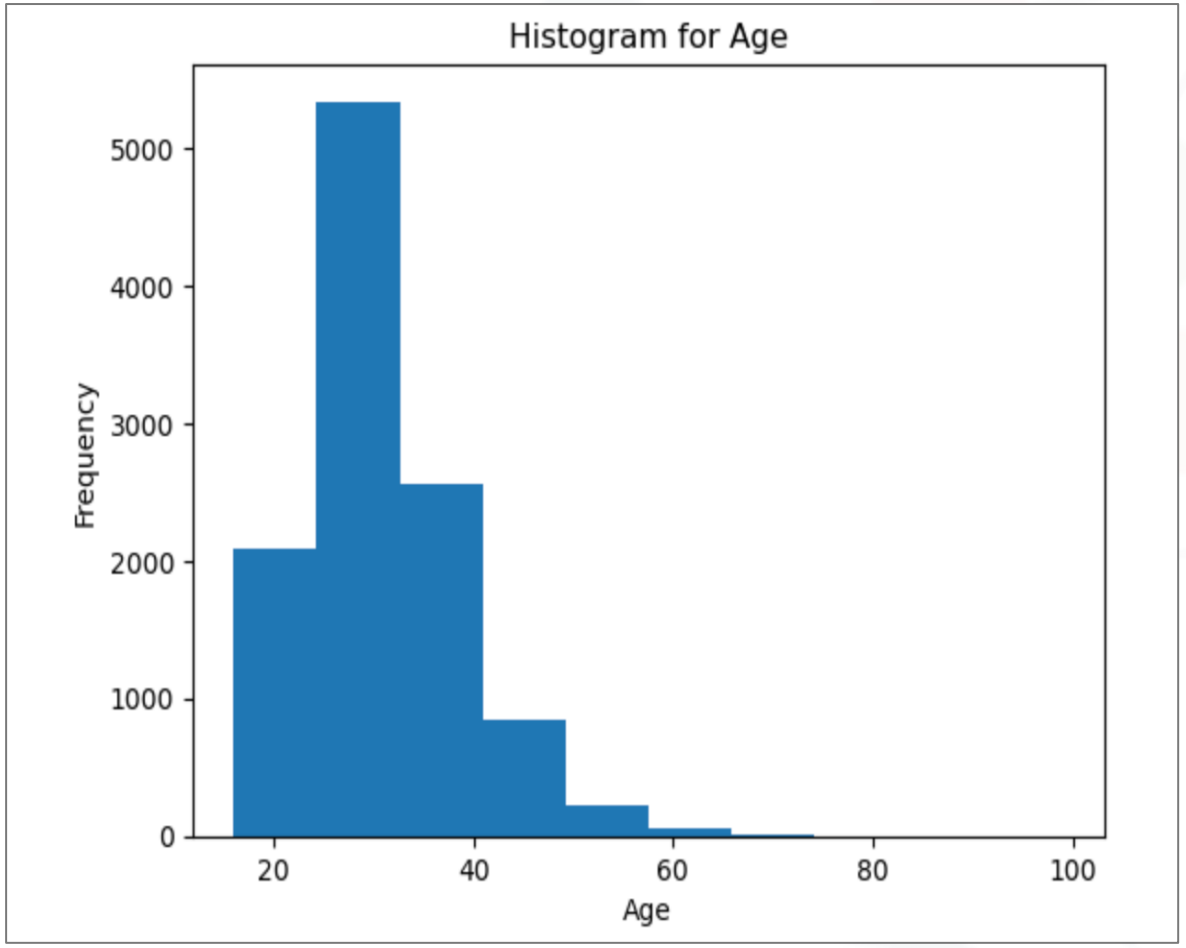


Fig - 9

Create a new dataframe by removing the outliers from the `ConvertedComp` column.

In [66]:

```
# your code goes here
#remove the outliers
mask = (df['ConvertedComp'] < (Q1 - 1.5 * IQR)) | (df['ConvertedComp'] > (Q3 + 1.5 * IQR))
import numpy as np
#change the outliers to 'na' to remove their numerical sata
df[mask] = np.nan
# check whether the data frame has changed
df['ConvertedComp'].mean()
```

Out[66]: 57785.980203205196

APPENDIX

Fig - 10

Correlation

Finding correlation

Find the correlation between `Age` and all other numerical columns.

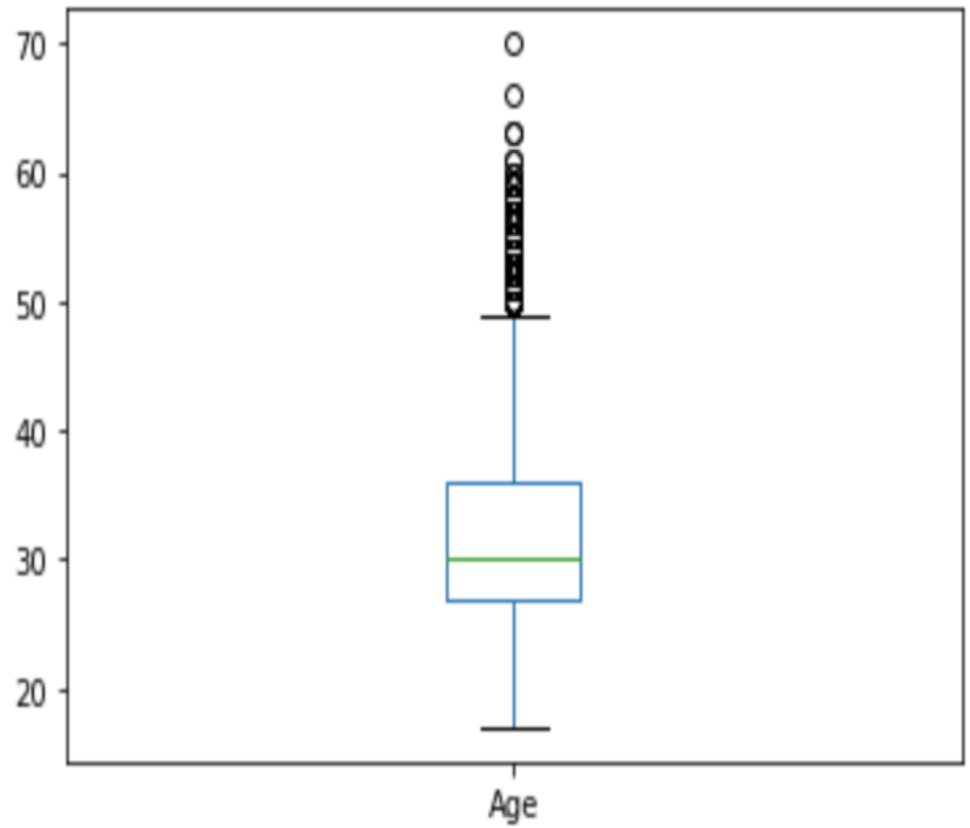
In [47]:
```
# your code goes here
df.corr()
```

Out[47]:

	Respondent	CompTotal	ConvertedComp	WorkWeekHrs	CodeRevHrs	Age
Respondent	1.000000	-0.019364	0.010878	-0.015275	0.002980	0.003950
CompTotal	-0.019364	1.000000	-0.063561	0.004975	0.017536	0.006371
ConvertedComp	0.010878	-0.063561	1.000000	0.034351	-0.088934	0.401821
WorkWeekHrs	-0.015275	0.004975	0.034351	1.000000	0.031963	0.037452
CodeRevHrs	0.002980	0.017536	-0.088934	0.031963	1.000000	-0.017961
Age	0.003950	0.006371	0.401821	0.037452	-0.017961	1.000000

Fig - 11

Out[7]: <AxesSubplot:>



APPENDIX

Fig - 12

Create a scatter plot of `Age` and `WorkWeekHrs.`

In [8]:
```
# your code goes here
df.plot(kind='scatter', x='Age', y='WorkWeekHrs', figsize=(10, 6), color='darkblue')
```

Out[8]: <AxesSubplot:xlabel='Age', ylabel='WorkWeekHrs'>

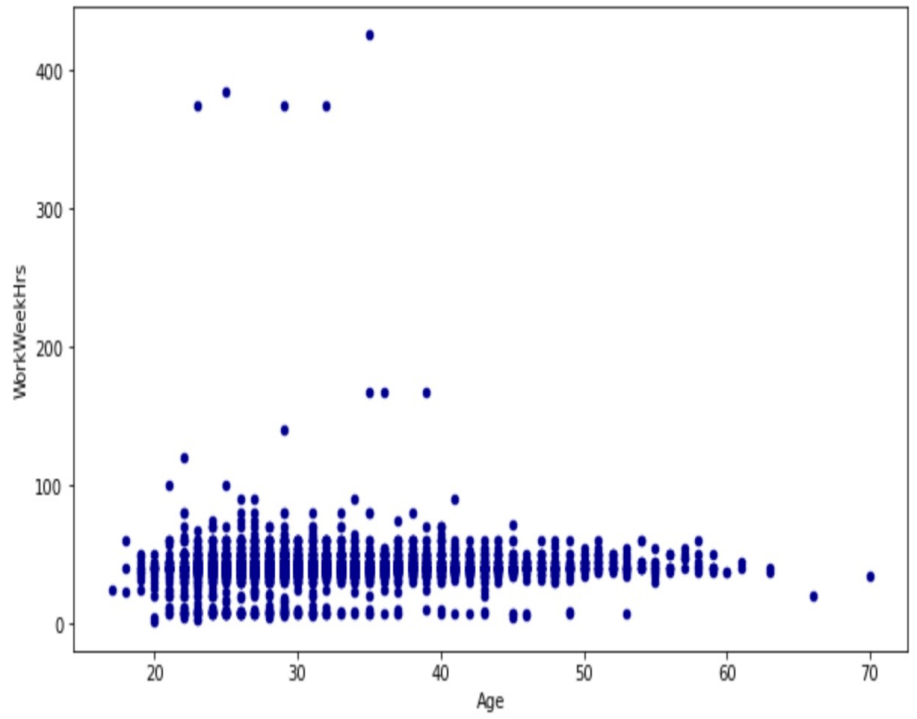
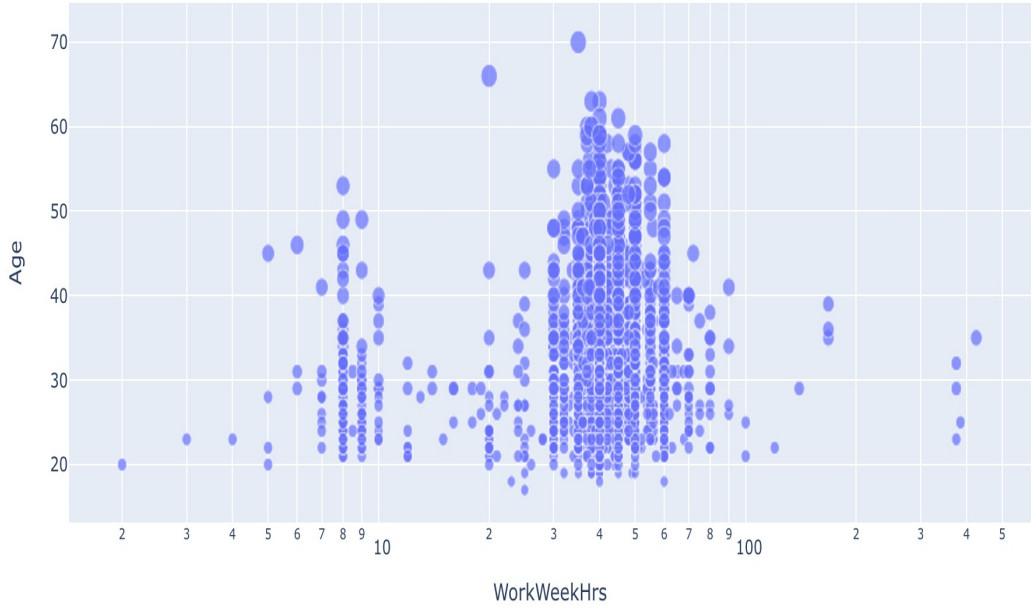


Fig - 13

Create a bubble plot of `WorkWeekHrs` and `Age`, use `Age` column as bubble size.

Hint: Use plotly.express to create a bubble chart

[8]:
```
# your code goes here
import plotly.express as px
fig = px.scatter(df, x="WorkWeekHrs", y="Age",
        size="Age", log_x=True, size_max=10)
fig.show()
```



APPENDIX

Fig - 14

Create a pie chart of the top 5 Country that respondents filled the survey . Display percentages of each database on the pie chart.

In [10]:
```
# your code goes here
df_pie = df['Country'].value_counts()
df_pie=df_pie.head(5)
df_pie.plot(kind='pie', figsize=(8,8), autopct='%1.0f%%')
```

Out[10]: <AxesSubplot:ylabel='Country'>

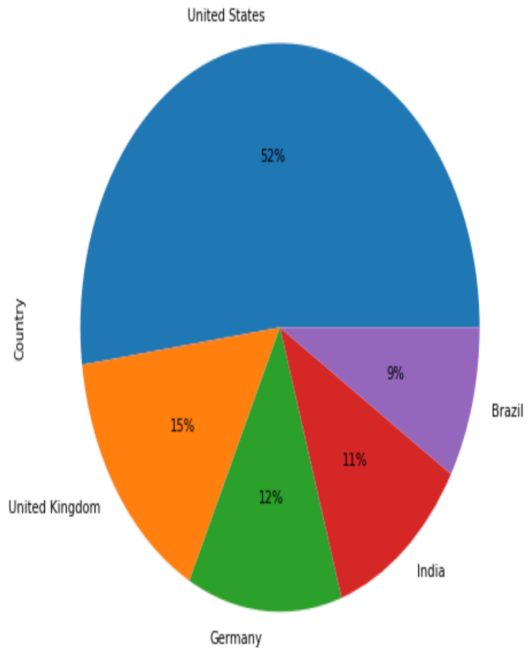
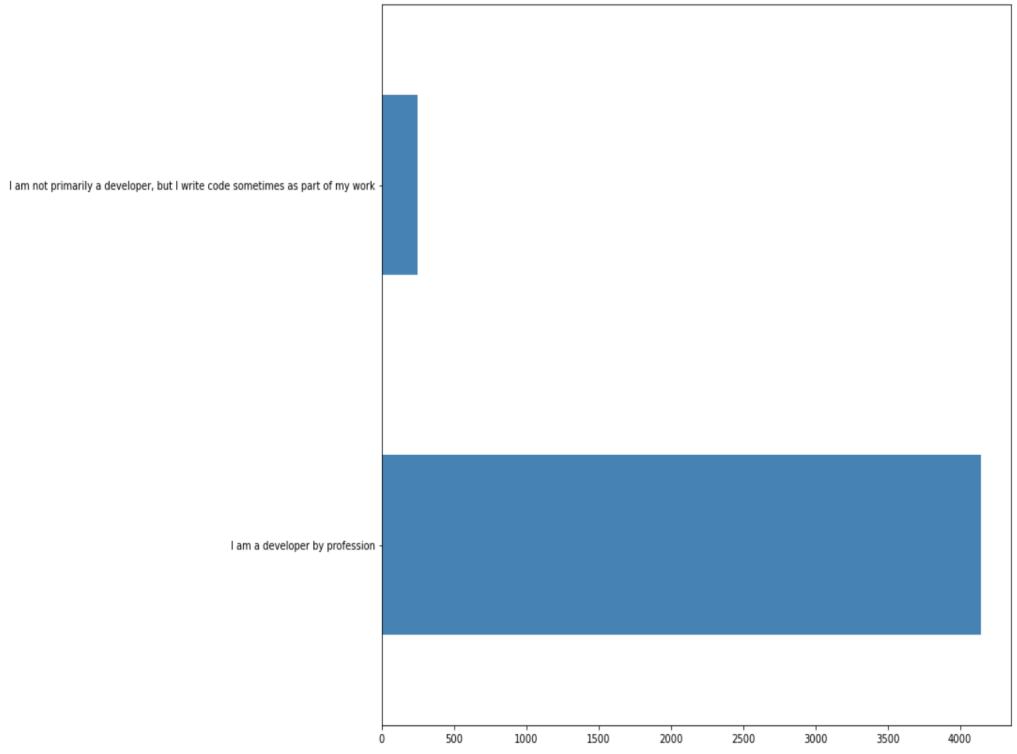


Fig - 15

Create a horizontal bar chart using column `MainBranch.`

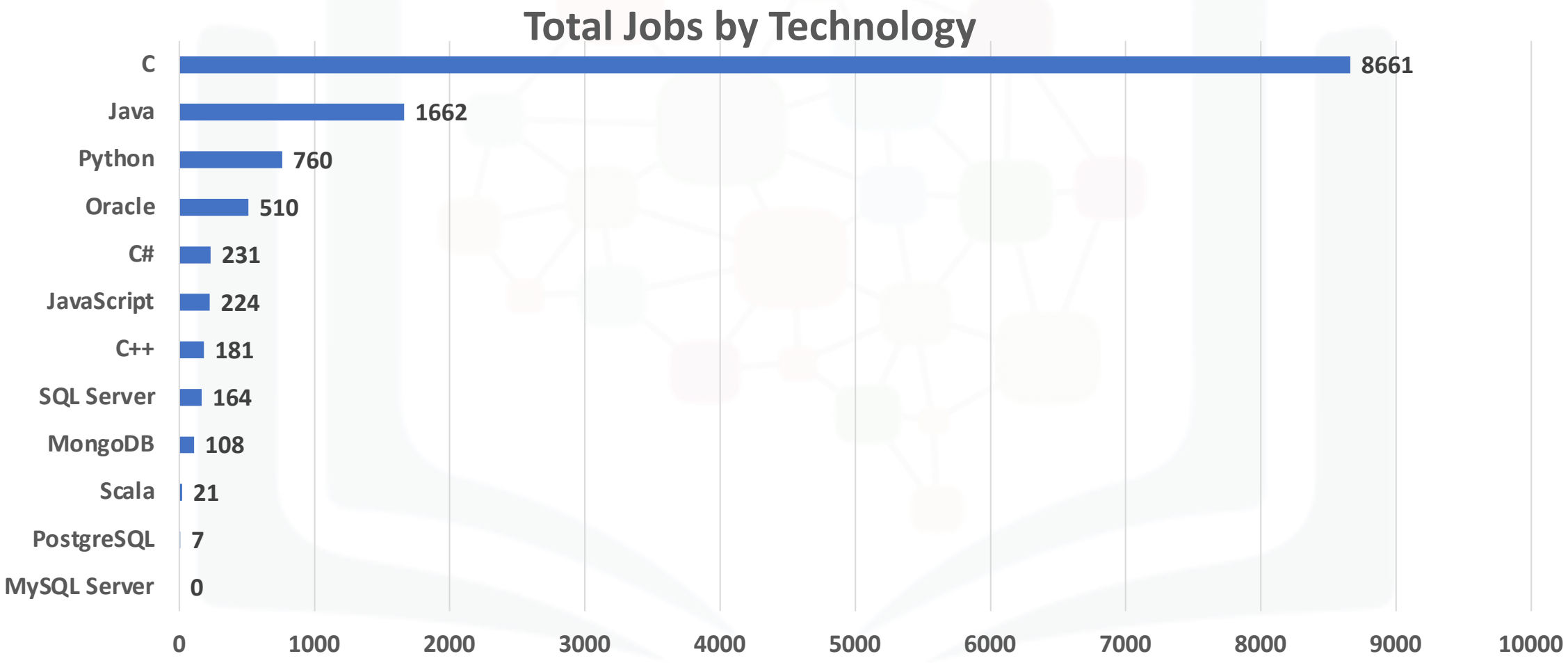
In [23]:
```
# your code goes here
df_mbranch = df['MainBranch'].value_counts()
df_mbranch
df_mbranch.plot(kind='barh', figsize=(12, 12), color='steelblue')
```

Out[23]: <AxesSubplot:>



JOB POSTINGS

Bar chart of the number of jobs postings by technology
(data collected in Module 1 using Job API in a file named “job-postings.xlsx”)



POPULAR LANGUAGES

Bar chart of the Annual Average Salary by Language
(data collected in Module 1 using web scraping in a file named “popular-languages.csv”)